Data Consistency and On-disk Write Buffering

Kai Shen

Dept. of Computer Science, University of Rochester

Consistency vs. Durability/Persistence

- After a system failure:
  - Consistency: ability to return to a consistent state – a state reachable by a series of legal operations:
    - in case of file systems: no leaking blocks, no dangling block pointers, all index block fields are in legal ranges
  - stronger consistent state - a state reachable by a series of legal operations that indeed occurred.
  - Durability/persistence: ability to return to a consistent state reachable by a series of legal operations that include those whose completions have been sent to users.

Consistency after Failure

- Fsck
- File system journaling
  - journal log before writes
  - journal clear after writes
- Physical/logical journaling
- Soft updates [McKusick and Ganger 1999]
  - failures allowed at any time without hurting consistency except leaks
  - still require ordering

On-disk Cache

- Disk cache
  - typically volatile memory
  - cache reads (is caching here really useful?)
  - buffer writes
- Writing buffering hurts data durability
- Writing buffering also hurts consistency
  - it breaks the journaling order constraints: journal log before actual writes; actual writes before journal clear

Consistency with Write Buffering

- Tagged command queueing
  - multiple writes can be outstanding (each tagged with an ID) and completion notification is sent back when a write is committed
  - allowing disk optimizations and maintaining order constraints
- Optional on both SCSI and ATA/SATA
  - commonly supported on SCSI but probably not so for ATA/SATA [McKusick 2006]

Other Alternatives

- 1, Synchronizing all writes
- 2, Flushing all buffered writes after each journal writes and before each journal clear
- 3, Synchronizing journal writes (and delay journal clears sufficiently)
- Any other ideas?
- Compare across them.
  - overhead of method 3 [Pearson 2008]

Linux Ext3

- At the absence of TCQ, no use of write synchronization [Pearson 2008]
- No support of write flushes/barriers up to Linux 2.6.11 [Nightingale et al. 2006]